文章编号: 1000-5862(2012)03-0288-04

基于模糊聚类广义回归神经网络的网络入侵研究

王　博[1], 彭玉涛[2*], 罗　超[2]

(1. 井冈山大学电子与信息工程学院, 江西 吉安 343009; 2. 井冈山大学现代教育技术中心, 江西 吉安 343009)

摘要: 采用结合模糊聚类和广义神经网络回归聚类分析的方法, 对5种网络入侵行为模式进行有效的聚类.首先用模糊 c 均值聚类算法将入侵数据分为 5 类, 再将聚类的结果中最靠近每类中心的样本作为广义神经网络的聚类训练样本进行数据训练, 训练输出的结果即为该个体所属的入侵类别. 实验结果表明: 新算法对网络入侵途径的分类精度更高, 可为预防网络入侵提供更可靠的数据支持.

关键词: 聚类算法; 模糊聚类; 广义回归神经网络; 网络入侵检测

中图分类号: TP 389.1　　文献标志码: A

0　引言

网络入侵是指试图破坏计算机和网络系统资源完整性、机密性或可用性的行为, 因此网络入侵行为检测是网络安全防御体系中非常重要的组成部分. 除了常规的基于主机[1]、基于网络的入侵检测技术[2]外, 近年来, 研究人员提出了一些新的入侵检测方法, 如文献[3]提出的基于归纳学习的入侵检测方法、文献[4]提出的基于免疫机理的入侵检测方法、文献[5-6]提出的基于数据挖掘的入侵检测方法等. 但是这几种方法都存在各自的局限性. 聚类算法是数据挖掘中常用的算法, 但由于网络入侵特征数据维数较高, 不同入侵类别的数据差别较小, 不少入侵模式并不能被准确分类. 因此, 本文采用模糊聚类和广义神经网络相结合的聚类算法对入侵数据进行更准确的分类, 为网络安全防御体系设计提供更精确的数据支持.

1　FCM 聚类算法

聚类算法是数据挖掘中经常使用的算法, 它将物理的或抽象的对象分为几个种群, 每个种群内部个体间具有较高的相似性, 不同群体间个体相似性较低. 模糊 c-均值聚类算法(Fuzzy C-Mean, FCM)是用隶属度确定每个元素属于某个类别程度的一种聚类算法, FCM 算法把 n 个数据向量 $\boldsymbol{x}_k$ 分为 c 个模糊类, 并求每类的聚类中心, 从而使模糊目标函数最小, 模糊聚类目标函数为

$$J=\sum_{i=1}^{n}\sum_{j=1}^{x}(u_{ij})^m\left\|\boldsymbol{x}_i-\boldsymbol{v}_j\right\|, \tag{1}$$

其中 u_{ij} 为个体 $\boldsymbol{x}_i$ 属于第 j 类的模糊隶属度, m 为模糊权重指数; $\boldsymbol{v}_j$ 为第 j 类的聚类中心, u_{ij} 和 $\boldsymbol{v}_j$ 的计算公式为

$$\boldsymbol{v}_i=\sum_{i=1}^{n}u_{ij}^m\boldsymbol{x}_i\Big/\sum_{i=1}^{n}u_{ij}^m, \tag{2}$$

$$u_{ij}=\begin{cases}\left[\sum\limits_{k=1}^{c}\dfrac{\left\|\boldsymbol{x}_i-\boldsymbol{v}_j\right\|^{\frac{2}{m-1}}}{\left\|\boldsymbol{x}_i-\boldsymbol{v}_k\right\|^{\frac{2}{m-1}}}\right]^{-1}, & \left\|\boldsymbol{x}_i-\boldsymbol{v}_k\right\|\neq 0,\\ 1, & \left\|\boldsymbol{x}_i-\boldsymbol{v}_k\right\|\neq 0\text{且}k=j,\\ 0, & \left\|\boldsymbol{x}_i-\boldsymbol{v}_k\right\|\neq 0\text{且}k\neq j.\end{cases} \tag{3}$$

FCM 聚类算法迭代过程如下: (i)给定类别书 c, 模糊权重指数 m; (ii)初始聚类中心 $\boldsymbol{v}$; (iii)根据(3)式计算幕后隶属度矩阵 $\boldsymbol{u}$; (iv)根据(2)式计算每类中心 $\boldsymbol{v}$; (v)根据(1)式计算模糊聚类目标值, 判断是否满足结束条件, 满足则算法终止, 否则返回(iii).

FCM 算法最终得到了模糊隶属度矩阵 $\boldsymbol{u}$, 个体根据隶属度矩阵每列最大元素位置判断个体所属类别.

收稿日期: 2011-12-20

基金项目: 国家自然科学基金(61063007)资助项目.

作者简介: 彭玉涛(1971-), 男, 广东兴宁人, 助理实验师, 主要从事软件应用和网络安全的研究.

2　广义回归神经网络

广义回归神经网络(GRNN)可归类于径向基神经网络, 其特点是具有较强的非线性映射功能, 它的网络可调性较强, 容错性和鲁棒性也很高, 非常适合用来分析和解决非线性问题. 同时, 广义回归神经网络在学习训练时有较好的收敛性, 学习曲线平滑不易震荡, 学习速率也较快, 因此, 在信号处理、专家系统、控制决策系统和结构分析等要处理不稳定数据的领域有广泛的应用.

2.1　广义回归神经网络模型

广义回归神经网络模型一般由输入层、模式层、求和层和输出层组成, 对应网络输入 $\boldsymbol{X}=[x_1, x_2,\cdots,x_n]^{\mathrm{T}}$, 其输出为 $\boldsymbol{Y}=[y_1,y_2,\cdots,y_k]^{\mathrm{T}}$.

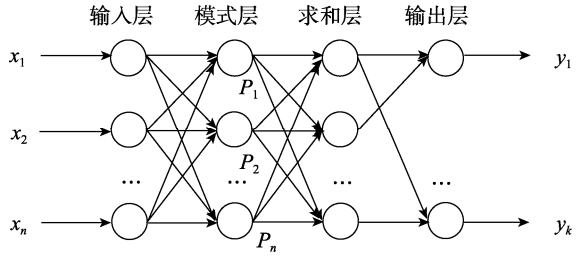

图 1　广义回归神经网络结构图

(i)输入层: 输入层有若干个神经元, 神经元的数量等于训练学习样本的维数, 输入层的数据通过神经元传递给模式层.

(ii)模式层: 模式层的神经元数目由学习样本的数量 n 决定, 该层的传递函数为

$$P_i=\exp\left[-\frac{(\boldsymbol{X}-\boldsymbol{X}_i)^{\mathrm{T}}(\boldsymbol{X}-\boldsymbol{X}_i)}{2\delta^2}\right],\quad i=1,2,\cdots,n.$$

(iii)求和层: 求和层的求和计算分为 2 类, 一类为算术求和, 另一类为加权求和. 当进行算术求和时, 求和公式为 $\sum_{i=1}^{m}\exp\left[-\frac{(\boldsymbol{X}-\boldsymbol{X}_i)^{\mathrm{T}}(\boldsymbol{X}-\boldsymbol{X}_i)}{2\delta^2}\right]$, 这时模式层的连接权值为 1, 并且传递函数为

$$S_D=\sum_{i=1}^{n}P_i.$$

当进行加权求和计算时, 求和公式为 $\sum_{i=1}^{m}\boldsymbol{Y}_i\exp\left[-\frac{(\boldsymbol{X}-\boldsymbol{X}_i)^{\mathrm{T}}(\boldsymbol{X}-\boldsymbol{X}_i)}{2\delta^2}\right]$, 此时传递函数为

$$S_{Nj}=\sum_{i=1}^{n}y_{ij}P_i,\ j=1,2,\cdots,k.$$

(iv)输出层: 输出层的神经元个数由学习训练样本输出的向量维数 k 来决定, 各神经元将求和层的输出相除, 神经元 j 的输出对应估计结果 $\hat{\boldsymbol{Y}}(\boldsymbol{X})$ 的第 j 个元素, 即

$$y_i=S_{Nj}/S_D,\quad j=1,2,\cdots,k.$$

2.2　广义回归神经网络的理论基础

非线性回归分析方法是广义回归神经网络的理论基础. 该方法的表明, 对于非独立变量 $\boldsymbol{Y}$, 要进行它相对于独立变量 $\boldsymbol{x}$ 的回归分析, 其实质是要计算出最大概率值 $\boldsymbol{y}$, 则可归纳出条件均值的计算公式为

$$\hat{\boldsymbol{Y}}=E(y\mid\boldsymbol{X})=\frac{\int_{-\infty}^{\infty}yf(\boldsymbol{X},y)\mathrm{d}y}{\int_{-\infty}^{\infty}f(\boldsymbol{X},y)\mathrm{d}y},\tag{4}$$

其中, $f(x,y)$ 为随机变量 x 和 y 的联合密度函数, $\boldsymbol{X}$ 为随机变量 x 的观测值. 同时, $\hat{\boldsymbol{Y}}$ 也可理解为当输入值为 $\boldsymbol{X}$ 时, $\boldsymbol{Y}$ 的预测输出值.

同时, 在已知样本数据集 $\{x_i,y_i\}_{i=1}^{n}$ 的情况下, 用 parzen 非参数估计来估算密度函数 $\hat{f}(\boldsymbol{X},\boldsymbol{y})$, 公式为

$$\hat{f}(\boldsymbol{X},\boldsymbol{y})=\frac{1}{n(2\pi)^{\frac{p+1}{2}}\sigma^{p+1}}\cdot\sum_{i=1}^{n}\exp\left[-\frac{(\boldsymbol{X}-\boldsymbol{X}_i)^{\mathrm{T}}(\boldsymbol{X}-\boldsymbol{X}_i)}{2\sigma^2}\right]\exp\left[-\frac{(\boldsymbol{X}-\boldsymbol{Y}_i)^2}{2\sigma^2}\right],$$

其中, X_i 为随机变量 x 和 y 的样本观测值, Y_i 为随机变量 y 的样本观测值, p 为随机变量, n 为样本的容量, x 为样本向量的维数; σ 称为光滑因子, 即高斯函数的宽度系数. 用上式的 $\hat{f}(\boldsymbol{X},y)$ 代替 $f(\boldsymbol{X},y)$, 则(4)式可表示为

$$\hat{\boldsymbol{Y}}(\boldsymbol{X})=\frac{\sum_{i=1}^{n}\exp\left[-\frac{(\boldsymbol{X}-\boldsymbol{X}_i)^{\mathrm{T}}(\boldsymbol{X}-\boldsymbol{X}_i)}{2\sigma^2}\right]\int_{-\infty}^{+\infty}y\exp\left[-\frac{(\boldsymbol{Y}-\boldsymbol{Y}_i)^2}{2\sigma^2}\right]\mathrm{d}y}{\sum_{i=1}^{n}\exp\left[-\frac{(\boldsymbol{X}-\boldsymbol{X}_i)^{\mathrm{T}}(\boldsymbol{X}-\boldsymbol{X}_i)}{2\sigma^2}\right]\int_{-\infty}^{+\infty}\exp\left[-\frac{(\boldsymbol{Y}-\boldsymbol{Y}_i)^2}{2\sigma^2}\right]\mathrm{d}y}.$$

由于 $\int_{-\infty}^{+\infty}z\mathrm{e}^{-z^2}\mathrm{d}z=0$, 则可计算出 $\hat{\boldsymbol{Y}}(\boldsymbol{X})$ 为

$$\hat{\boldsymbol{Y}}(\boldsymbol{X})=\frac{\sum_{i=1}^{n}Y_i\exp\left[-\frac{(\boldsymbol{X}-\boldsymbol{X}_i)^{\mathrm{T}}(\boldsymbol{X}-\boldsymbol{X}_i)}{2\sigma^2}\right]}{\sum_{i=1}^{n}\exp\left[-\frac{(\boldsymbol{X}-\boldsymbol{X}_i)^{\mathrm{T}}(\boldsymbol{X}-\boldsymbol{X}_i)}{2\sigma^2}\right]},$$

可以看出, $\hat{\boldsymbol{Y}}(\boldsymbol{X})$ 本质是 Y_i 的加权均值, $\boldsymbol{Y}_i$ 的权重因

子为样本 $\boldsymbol{X}_i$ 与 X 的距离平方指数. $\hat{\boldsymbol{Y}}(\boldsymbol{X})$ 的误差与光滑因子 σ 有非常紧密的联系, 可分 2 种情况讨论: (i)当 σ 取值非常大时, $\hat{Y}(X)$ 接近样本变量均值; (ii)当 σ 趋近于 0 时, $\hat{Y}(X)$ 接近于学习训练样本.

由于网络较差的泛化力, 只有当 σ 取得合适的值时, $\hat{Y}(X)$ 才会有更理想的取值. 同样的, 当需要被预测的点被训练样本所包含时, 预测值将非常接近学习样本的因变量的值, 则将取得较好的预测值, 反之, 预测误差会比较大, 预测效果也较差. 此时, 可以考虑对距离较近的预测点的因变量赋予更大的权值.

3 模糊聚类和广义回归神经网络的结合算法

由于网络入侵特征数据维数较多, 不同入侵类别间的数据差别较小, 如果简单采用模糊聚类进行数据挖掘, 则会造成较多入侵模式不能被准确分类的后果. 本文采用结合模糊聚类和广义神经网络回归的聚类算法对入侵数据进行分类. 算法的流程如图 2 所示.

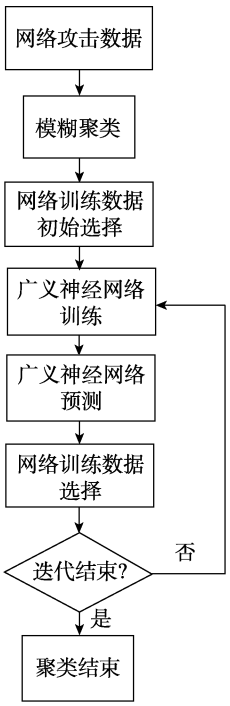

图 2 算法流程图

3.1 模糊聚类模块

用模糊聚类算法把侵入数据分为 n 类, 并将得到每类的聚类中心和个体模糊隶属度矩阵 $\boldsymbol{u}$. 此模块中应使用模糊聚类函数 $fcm()$.

3.2 网络训练初始数据选择模块

根据模糊聚类的结果选择最靠近每类中心的样本作为广义神经网络聚类训练样本. 首先, 求每类的类内均值 $mean_i(i=1,2,\cdots,n)$, 然后求解每类中所有样本 $\boldsymbol{X}$ 到中心值 $mean_i(i=1,2,\cdots,n)$ 的距离矩阵 $\boldsymbol{ecent}_i\ (i=1,2,\cdots,n)$, 从矩阵 $\boldsymbol{ecent}_i\ (i=1,2,\cdots,n)$ 中选择距离最小的 m 个样本作为一组, 设定其对应的网络输出为 i. 这样就得到了 $n\times m$ 组训练数据, 其输入数据为网络入侵特征数据, 输出数据为该入侵行为所属入侵类别.

3.3 广义神经网络训练模块

用训练数据训练广义神经网络. 此模块应使用广义神经网络训练函数 $newgrnn()$.

3.4 广义神经网络预测模块

用训练好的网络预测所有输入样本数据 $\boldsymbol{X}$ 的输出序列 $\boldsymbol{Y}$, 此模块应使用预测函数 $sim()$.

3.5 网络训练数据选择模块

根据预测输出把入侵数据重新分为 n 类, 并从中找出最靠近每类中心值的样本作为训练样本. 首先按照网络预测输出序列 $\boldsymbol{Y}$ 把样本数据 $\boldsymbol{X}$ 分为 n 类, 然后求出每类内所有样本平均值 $mean_i(i=1,2,\cdots,n)$, 求解出所有样本 $\boldsymbol{X}$ 到中心值 $mean_i(i=1,2,\cdots,n)$ 的矩阵 $\boldsymbol{ecent}_i\ (i=1,2,\cdots,n)$, 从距离矩阵 $\boldsymbol{ecent}_i\ (i=1,2,\cdots,n)$ 选择距离最小的 m 个样本作为一组, 设定其对应的网络输出为 i. 这样再次得到了 $n\times m$ 组网络训练数据, 其输入数据为网络入侵提取数据, 输出数据该个体所属入侵类别.

本案例的数据来自 5 种网络入侵数据, 算法的目的是能够对这 5 种入侵数据进行有效聚类.

4 实验结果分析

根据 FCM 聚类算法和广义神经网络原理, 在 MATLAB 中编程实现基于广义神经网络的聚类算法, 用神经网络对 4 000 组入侵数据按照本文的算法进行聚类, 以达到分类 5 种网络入侵数据的目的.

算法反复计算 10 次, 最后得到的聚类结果及 2 种聚类方法的比较如表1 所示, 其中, 每行均表示聚类算法得到每类样本在实际各入侵类别中的分布数量.

表 1　聚类结果比较

入侵类别	模糊聚类					广义神经网络模糊聚类				
	聚类结果 1	聚类结果 2	聚类结果 3	聚类结果 4	聚类结果 5	聚类结果 1	聚类结果 2	聚类结果 3	聚类结果 4	聚类结果 5
实际入侵类别 1	6	1 211	332	14	0	31	1 171	345	16	0
实际入侵类别 2	0	0	0	0	2 097	0	0	0	0	2 097
实际入侵类别 3	0	0	0	0	130	39	2	1	1	87
实际入侵类别 4	0	0	0	0	658	658	0	0	0	0
实际入侵类别 5	0	0	0	0	52	0	0	0	0	0

从表1 中可以看出, 对于网络入侵数据, 模糊聚类没有实现对数据的有效分类, 聚类结果没有把类别 2 到类别 5 样本区分开来, 但广义神经网络模糊聚类对其进行了有效的分类.

5　结论

本文结合了模糊聚类的无导师聚类和广义神经网络的有导师学习功能完成了对未知网络入侵数据的聚类, 广义神经网络所起的作用为训练后分类所有入侵样本, 实验的结果表明, 使用本文的算法的聚类结果比简单的模糊聚类结果要更准确, 更能有效的对网络入侵数据进行聚类, 因此可为网络的安全机制的设计提供更为可靠的数据支持.

6　参考文献

[1] 张艳艳, 彭新光. 虚拟健壮主机入侵检测的实验研究 [J]. 计算机应用与软件, 2010, 27(4): 130-132.

[2] 李亮超. 基于校园网的网络入侵检测系统的研究与实现 [J]. 微型电脑应用, 2010, 26(3): 61-64.

[3] 刘培顺, 王学芳. 入侵检测中的归纳学习方法 [J]. 计算机工程, 2006, 32(16): 125-162.

[4] 方贤进, 李龙澍, 钱海. 基于人工免疫的网络入侵检测中疫苗算子的作用研究 [J]. 计算机科学, 2010, 37(1): 239-242.

[5] 李涛. 基于数据挖掘技术的自适应入侵检测系统模型 [J]. 计算机工程与设计, 2010, 31(6): 1209-1211.

[6] 傅德胜, 周舒, 郭萍. 基于数据挖掘的分布式网络入侵检测系统设计及实现 [J]. 计算机科学, 2009, 36(3): 103-105.

[7] 王博. 基于 BP 网络的一种改进算法及仿真 [J]. 井冈山学院学报: 自然科学版, 2009, 30(10): 37-39.

[8] 王博, 李冬妹, 罗超. 基于小波神经网络的非线性系统工程安全性评价研究 [J]. 井冈山大学学报: 自然科学版, 2010, 31(3): 78-82.

[9] 杨群, 蔡乐才, 莫再群, 等. 基于组合智能的网络入侵检测模型 [J]. 微计算机信息, 2010, 26(2): 89-91.

[10] 张桂林, 柯永振, 李智超, 等. 一种入侵预测系统的建模与仿真研究 [J]. 系统仿真学报, 2010, 22(7): 1796-1799.

[11] 曹从军, 孙静. 基于广义回归神经网络的数码打样色彩空间转换方法 [J]. 计算机应用, 2010, 30(8): 2108-2110.

[12] 王小军. 模糊聚类在入侵检测中的应用研究 [D]. 南京: 南京师范大学, 2006

[13] 李辉, 蔡敏, 李宇, 等. 基于自适应粒子群优化算法的神经网络的优化研究 [J]. 江西师范大学学报: 自然科学版, 2010, 34(6): 632-635.

[14] 江琴, 程树森, 杜后发. BP 神经网络在烧结矿技术经济指标的应用 [J]. 江西师范大学学报: 自然科学版, 2009, 33(4): 414-416.

The Clusting Research for Net Attack Based on Fuzzy Clustering and GRNN

WANG Bo[1], PENG Yu-tao[2*], LUO Chao[2]

(1. School of Electronics and Information Engineering, Jinggangshan University, Ji′an Jiangxi 343009, China;
2.Center of Moderm Education Technology, Jingganshan University, Ji′an Jiangxi 343009, China)

Abstract: The methods of fuzzy clustering and GRNN clustering, which this article used, to divide net attack behavior into five categories. First, use fuzzy c-mean clustering arithmetic divide attack data into five categories; then, from the result of the clustering, these samples, which closest to the every clustering center, is going to data training as GRNN clustering training simples, the output result of training. is the attack category of this data belonged to. This method has higher classification accuracy in how to divide net attack methods into, and net attack instrusion prevention with more reliable data support.

Key words: clustering; fuzzy clustering; GRNN; net attack testing

(责任编辑: 冉小晓)